УДК 004.9:371.261

Н.Б. Шаховська, І.Б. Швороб
Національний університет «Львівська політехніка», Україна
вул. С. Бандери, 12, м. Львів, 79013

МЕТОД ПОБУДОВИ ТЕКСТОВОГО ШАБЛОНУ ДЛЯ ЕКСТРАКЦІЇ ІНФОРМАЦІЇ ЗІ СЛАБОСТРУКТУРОВАНИХ ДАНИХ

N.B. Shakhovska, I.B. Shvorob
Lviv Polytechnic National University, Ukraine
12, S. Bandery str., Lviv, 79013

METHOD OF CONSTRUCTING A TAXT TEMPLATE FOR EXTRUCTING INFORMATION FROM SEMISTRUCTURED DATA

80% світових даних є неструктурованими або слабоструктурованими. У зв’язку з цим, актуальною є проблема екстракції інформації та її подальше збереження у зручній для опрацювання формі. Для зручності екстракції даних у роботі запропоновано використання текстових шаблонів на основі словника ключових слів. Основною метою є розроблення методу виділення складових елементів для побудови текстового шаблону, а також розроблення методу кластеризації текстового шаблону. Проведено аналіз розроблених методів на прикладі роботи бібліотечної системи.

Ключові слова: слабоструктуровани дані, екстранкція даних, текстові шаблони, методи кластеризації.

80% of world data is unstructured or semistructured. In this regard, the main task is the problem of extraction of information and its further preservation in a form suitable for processing. For the convenience of data extraction, we suggest using text templates based on the dictionary of keywords. The main goal is to develop a method for selecting component elements for constructing a text template, as well as developing a method for clustering a text template. The analysis of the developed methods on the example of work of the library system is carried out.

Keywords: semistructured data, data extruction, text templates, methods of clasterisation.

Вступ

За даними досліджень, 80% світових даних є неструктурованими або слабоструктурованими. У зв’язку з цим, актуальною є проблема екстракції інформації та її подальше збереження у зручній для опрацювання формі.

Слабкоструктурованими даними є будь-які проміжні дані між структурованими й неструктурованими. Такі дані мають певні особливості. По-перше, структура даних може бути неповною, недовизначеною. По-друге, значення скалярних даних представлені у вигляді текстової інформації. По-третє, виникає проблема визначення приналежності даних, тому що не завжди можна однозначно судити про коректність оброблюваного документа.

Однією з основних проблем у роботі зі слабоструктурованими даними є різноманітність даних. Питання різноманітності даних в інформаційних системах є складним, воно також включає в себе такі області, як групування даних за певними характеристиками без послідовного перекриття множин даних.

Для вирішення цих питань у роботі пропонується застосування підходу побудови текстового шаблону на основі заданих ключових слів. Такий підхід допоможе відсіювати зайву інформацію та отримувати необхідні дані, уникаючи дублювань.

Вхідною інформацією для віднесення тексту до текстового шаблону є текстовий файл будь-якого формату зі слабоструктурованим текстом. З файлу необхідно визначити базові характеристики на основі сформованого словника ключових слів (маркерів).

У даній роботі основною метою є розроблення методу виділення складових елементів для побудови текстового шаблону, а також розроблення методу кластеризації текстового шаблону. Проведено аналіз розроблених методів на прикладі роботи бібліотечної системи.

1. Розроблення методу виділення складових елементів для побудови текстового шаблону

Текстовий шаблон складається з послідовності речень $A_1, A_2, ..., A_l$ та утворює кортеж $T = (A_1, A_2, ..., A_l)$, а речення $A_i, i = \overline{1,k}$ – з послідовності слів $a_{ij}, i = \overline{1,l}, j = \overline{1,n}$, яка, у свою чергу, зображується кортежем $r_i = (a_{i1}, a_{i2}, ..., a_{il})$. Позначимо через $|a_{ij}|$ довжину слова a_{ij}. Зміст (семантику) тексту T позначимо $S(T)$.

Введемо множину ключових слів (маркерів) $Key = \{key_1, key_2, ..., key_m\}$ шаблону, які містяться у досліджуваних текстах. У реченні $r = (a_1, a_2, ..., a_l)$ знаходять ключове слово $a_p (a_p \in Key)$.

Текстовий шаблон – це неструктурований або напівструктурований файл, який складається з послідовності речень, а речення – з послідовності слів. Зі всієї множини слів у документі вибираються тільки ті, що мають змістовне наповнення, тобто формується база даних «Ключові слова».

Метод формування бази даних «ключові слова» (функція f_2) передбачає наступні етапи:

Етап 1. Слабоструктурована текстова інформація розбивається на речення та слова.

Етап 2. Відкидаються слова, що містять менше трьох символів.

Етап 3. Здійснюється класифікація слів шляхом видалення з загального списку слів, які містяться в базі даних «Стоп-слова» та неінформативних слів і словосполучень.

Етап 4. Формується загальний список слів у документі, при цьому зберігається інформація про їх форматування та місце в тексті.

Етап 5. Загальний список слів модифікується в процесі стеммінгу, тобто відкидаючи закінчення слів, ми також видаляємо однакові слова з бази даних, але збільшуємо значення, що відповідає за кількість вживань цього слова в тексті, а ваги, що були попередньо присвоєні цим словам, додаються. Таким чином утворюється база даних «Ключові слова тексту».

Користувач може вносити свої ключові слова і визначати їх вагу, таким чином спрямовуючи систему на виділення інформації, яка пов’язана з введеними ключовими словами.

До бази даних «Стоп-слова» входять службові частини мови, тобто сполучники, а також займенники, вставні слова та інше.

Метод виділення складових текстового документа також базується на понятті ваги речення і розрахований на опрацювання наукових статей. Основу аналітичного етапу в цій моделі складає процедура призначення вагових коефіцієнтів для кожного блоку тексту відповідно до таких характеристик, як:

- розташування цього блоку в оригіналі;
- частота появи в тексті;
- частота використання в ключових реченнях;
- показники статистичної значущості.

Сума індивідуальних ваг слів та речення, як правило, визначена після додаткової модифікації відповідно до спеціальних параметрів налаштування, пов'язаних з кожною вагою, дає загальну вагу речення U :

$$Weight(U) = WordsWeighn(U) + 10 * Place(U) + 10 * Format(U) \quad (1)$$

Отже, найважливішими факторами для ваги речення вважатимемо формат та розташування.

Для формування реферату виділяються речення з основної частини.

Основна частина, у свою чергу, ділиться на фрагменти за підрозділами та розділами, введеними авторами. Вважається, що речення, що з'являються у вступній частині та висновках, мають більше інформативне значення, ніж речення із середини тексту [2].

У першу чергу, введемо поняття ваги речення. Для цього формалізуємо елементи (1).

Коефіцієнт розташування визначається як:

$$Place(U) = \begin{cases} 0, \left(\frac{n}{n_{count}} > 0{,}9\right) \vee \left(\frac{n}{n_{count}} < 0{,}1\right) \\ 1, \left(0{,}1 \le \frac{n}{n_{count}} < 0{,}3\right) \vee \left(0{,}7 < \frac{n}{n_{count}} < 0{,}9\right) \\ 2, \left(0{,}3 \le \frac{n}{n_{count}} \le 0{,}7\right) \end{cases}, \quad (2)$$

де n – номер речення, а n_{count} – загальна кількість речень у документі. Початок та кінець тексту оцінюються меншим значенням (бо це переважно вступ та висновок) 0-1, а середина – 2. Також, якщо у документі є анотація, яка переважно знаходиться між заголовком і вступом, то цьому фрагменту тексту присвоюється $Place(U) = 4$.

Коефіцієнт форматування речення U визначається як:

$$Format(U) = \begin{cases} 0, \text{вирівнювання зліва або справа} \\ 1, \text{вирівнювання поширині} \\ 2, \text{вирівнювання по центру} \end{cases} \quad (3)$$

Речення, що мають вирівнювання зліва або справа вважаються менш значущими, бо це переважно зазначення автора, УДК, дати та іншої додаткової інформації. Основна частина тексту зазвичай має вирівнювання за шириною, тому речення з таким форматуванням оцінюються більше, а речення, вирівнювання яких є посередині: це, як правило, заголовки або підзаголовки, тому дістають найвищу оцінку важливості.

Коефіцієнт $WordsWeighn(U)$ визначається як середня вага слова у реченні (сума ваг усіх ключових слів, що входять до речення, поділена на кількість ключових слів у реченні), таким чином довгі речення не будуть мати переваги над короткими.

Вага слова Q визначається за формулою:

$$Weight(Q) = Frequency(Q) + Place(Q) + Format(Q) + User(Q) \quad (4)$$

Частотний коефіцієнт $Frequency(Q)$ (frequency – частота) – відношення числа входження деякого слова (*word*) до загальної кількості слів (*words*) документа. Таким чином, оцінюється важливість слова в межах окремого документа:

$$Frequency(Q) = \frac{word}{words}. \quad (5)$$

Коефіцієнт розташування $Place(Q)$ визначається як функція належності до речення, де зустрічається слово, однієї з ключових фраз: «Ключові слова:» або «Ключевые слова:». Якщо така фраза зустрілась, то коефіцієнт розташування рівний 5, якщо ні – 0.

Коефіцієнт форматування слова $Format(Q)$ визначається залежно від того чи слово виділене жирним, курсивом чи підкреслене. Якщо слово зовсім не відформатоване, то коефіцієнт дорівнює 0, якщо одним форматом, то – 1, якщо двома, то – 2, якщо трьома, то – 3.

Показник $User(Q)$ формується на основі оцінювання слова користувачем, $User(Q) \in [0..10]$.

Вагові коефіцієнти, використані у формулі (1), отримані емпірично. У роботі ставилася задача не точного визначення їх значень, а встановлення ваги певних адитивних параметрів. Тому для цих коефіцієнтів важливим є порядок числа, а не його значення.

Результатом методу виділення складових текстового документа є *вектор*, у якому для певних характеристик тексту використовуються бінарні ознаки, а для ключових слів – ваги.

2. Розроблення методу кластеризації текстових шаблонів

Кластеризація – це автоматичне розбиття елементів деякої множини на групи. Кластеризацію проводитимемо модифікованим методом к-найближчих сусідів. Удосконалення вказаного методу здійснено з тією метою, що основним недоліком цього методу є залежність якості розбиття від кількості заданих користувачем кластерів. У випадку розбиття наукових публікацій кількість кластерів наперед невідома.

Існуючі методи кластеризації мають ряд обмежень для кластеризації наукових публікацій на наукові школи. Тому удосконалено метод k-середніх.

Алгоритм кластеризації – це відображення $f : X \to \{X_i\}$, яке будь-якому тексту $x \in X$ ставить у відповідність мітку кластера $X' \in \{X_i\}$.

Основна мета кластерного аналізу – знаходження груп схожих об'єктів у вибірці. Типи вхідних даних для кластерного аналізу:

- опис об'єктів на основі характерних ознак;
- матриця відстаней між об'єктами;
- матриця подібності між об'єктами.

Один з найбільших недоліків методу k-середніх і йому подібних полягає у тому, що вимагає попереднього вказання кількості кластерів, і від цієї кількості сильно залежить кластерне рішення. Тому в роботі вирішено модифікувати цей метод.

Модифікований метод k-середніх полягає у виконанні таких етапів:

1.Задаємо кількість кластерів *k*, $N >= k >= 2$, де *N* – кількість публікацій.

На вхід методу отримуємо множину ЕД, поданих у вигляді числових векторів.

Оскільки ознаки кластеризації (автор, наукова установа, назва, ключові слова) невпорядковані, то використовуватимемо метрику *d* ізольованих точок:

$$l(X.x, Y.x) = \begin{cases} 1, X.x = Y.x \\ 0, X.x \neq Y.x \end{cases},$$

$$d(X,Y) = \sum_i^p l(X.A_i, Y.A_i) + \sum_i^r l(X.D_i, Y.D_i) + \sum_i^w l(X.B_i, Y.B_i) + l(X.C, Y.C),$$

де функція l повертає 1, якщо обидва її параметри мають однакові значення, та 0 в іншому випадку, X, Y – електронні версії текстів наукових публікацій, p – кількість авторів в текстах публікацій X, Y, r – сумарна кількість ключових слів, w – сумарна кількість наукових установ, $X.A_i$ – значення автора x_i публікації X, $X.C$ – значення назви C наукової статті X.

$x \in R^n$ називається ізольованою точкою множини E, якщо будь-який окіл цієї точки не містить інших точок E, крім самої x:

$$d(x,y) = \begin{cases} 0, x = y \\ 1, x \neq y \end{cases}.$$

Будь-яка точка дотику множини E є або граничною, або ізольованою.

2. Обираємо k об’єктів, які вважатимемо центрами відповідних кластерів (центроїдами). Покласти номер кроку $s = 0$.

3. Формуємо вектор центроїдів $\left\langle cx_1^s, cx_2^s, ..., cx_k^s \right\rangle$ (центрів ваги).

Для кожного об’єкта знаходимо відстань до усіх центроїдів. Для знаходження відстані використовуємо Евклідову метрику.

4. Шукається матриця відстаней до центроїдів кластерів і формуємо кластери S_i, $i = \overline{1,k}$.

$$\min \left[\sum_{j=1}^{k} \sum_{i}^{N} \left\| x_i - cx_j \right\|^2 \right],$$

де N – кількість публікацій, cx_j – центроїд кластера з номером j.

Після розрахунку матриці відстаней шукаються *сильні зв’язки* об'єкта з кластером.

Сильним названо зв’язок між об’єктами X та X_i, якщо значення відстані назв публікацій менше, ніж третина від максимальної відстані серед усіх назв публікацій:

$$d_s(X, X_i) \leq \frac{\max\left(d(X, X_1), ..., d(X, X_N)\right)}{3}.$$

5. Шукаємо вартість розбиття:

$$Cost = \sum_{i=1}^{k} \sum_{j=1}^{|S_i|} d_{ij} d_s(x_j, cx_i),$$

де k – кількість кластерів, $|S_i|$ – кількість об’єктів у кластері S_i, d_{ij} – відстань до центру кластера i.

6. Шукаємо нові центроїди кластерів:

$$cx_i = \frac{1}{|S_i|} \sum_{x_j \in S_i} x_j .$$

Якщо $\left\| CX^s \right\| \neq \left\| CX \right\|$, то $s = s + 1$. Перейти на крок 3.

7. Якщо $N > k$ і $Cost$ не задовольняє умовам локального оптимуму, k=k+1 і перейти на крок 3.

3. Апробація методу кластеризації текстових шаблонів

Апробацію методу кластеризації текстових шаблонів здійснено на прикладі бібліотечних систем.

Система кластеризації наукових публікацій може бути використана електронними бібліотеками для попереднього аналізу текстів та їх рубрикування. Ця задача вироджується у задачу класифікації, оскільки кількість кластерів, їх назви та ознаки (ключові слова) є відомими.

Є такі первинні налаштування системи для попереднього рубрикування:

idClusters	Theme
1	Database
2	Computer science
3	Programming
4	Network
5	System analysis

Рис.1. Дані таблиці «Cluster»

idWords	Word
1	diagram
2	data
3	document
4	analyse
5	protocol
6	algorithm
7	system
8	table
9	code
10	connection
11	module
12	library
13	кластер
14	обчислення

Рис. 2. Дані таблиці «KeyWord»

id	Clusters_id	Words_id
1	1	2
2	1	8
3	1	7
4	1	10
5	1	13
6	1	16
7	1	14
8	2	2
9	2	5
10	2	7
11	2	9
12	2	10
13	2	18
14	3	6
15	3	7
16	3	11
17	3	12

Рис. 3. Дані таблиці «Classification»

Для тестування роботи системи опрацьовано 134 файли наукових публікацій, поданих у форматі MS Word. «Правильна» рубрика текстових документів відома наперед і встановлена експертно.

Проаналізуємо якість рубрикації (*TP* (*true positive*) – кількість ЕД, правильно віднесених до категорії; *FP* (*false positive*) – помилка другого роду – кількість ЕД, неправильно віднесених до категорії; *FN* (*false negative*) – помилка першого роду – кількість ЕД, які неправильно відкинуті; *TN* (*true negative*) – кількість ЕД, які правильно відкинуті):

Середнє нормоване значення правильно рубрикованих документів становить 94 %. Середнє нормоване значення неправильно віднесених до категорії документів становить 8%, оскільки, як видно з рис. 2, майже усі класи мають спільні ключові слова. Середнє нормоване значення неправильно відкинутих документів становить 6% і середнє нормоване значення правильно відкинутих документів становить 44%.

Таблиця 1. Результати аналізу якості кластеризації

Клас	nTP	nFP	nFN	nTN
Database	93%	11%	7%	33%
computer science	93%	13%	7%	25%
Programming	96%	2%	4%	50%
Network	94%	6%	6%	60%
system analysis	93%	7%	7%	50%

Далі проаналізовано залежність якості кластеризації від обсягу класів. У навчальній вибірці присутні класи з великою кількість представників і класи з малою кількістю представників (таблиця 2). Є класи, що містять більше, ніж 50% статей, інші містять тільки 2%.

Таблиця 2. Кількість статей за класами

Клас	К-сть	%
Database	21	16%
computer science	74	55%
Programming	31	23%
Network	5	4%
system analysis	3	2%

При цьому зрозуміло, що чим більшою є «загальність» рубрики, тим важче її кластеризувати. Рис. 4 експериментально підтверджує цю гіпотезу.

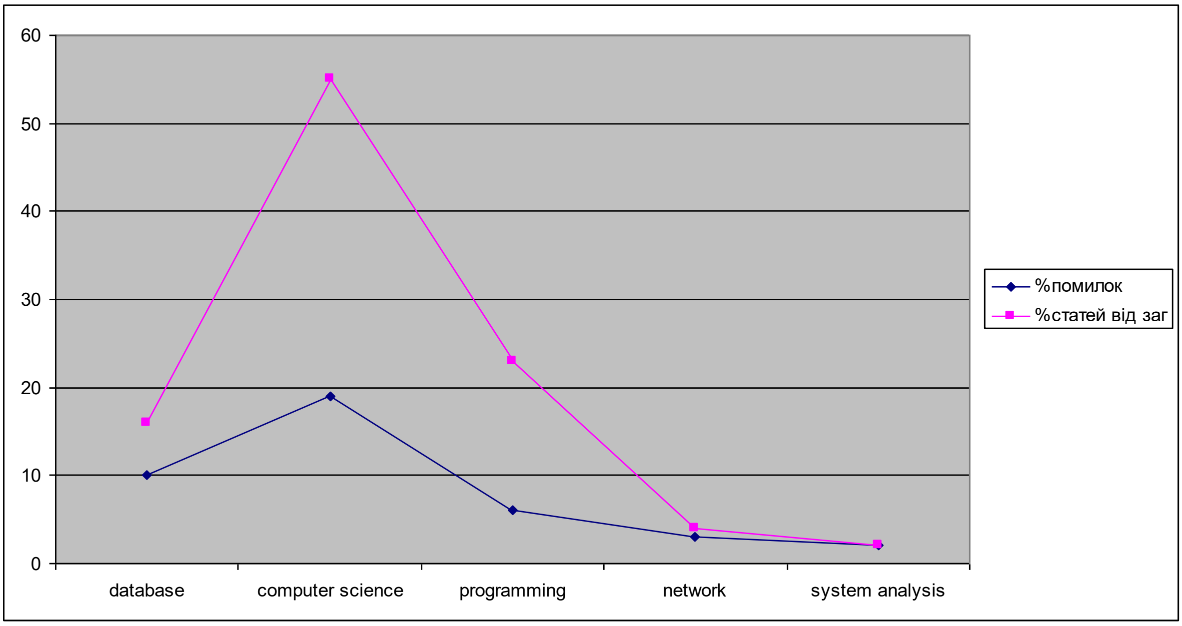


Рис. 4. Залежність помилки першого роду від обсягу вибірки

Кореляційний момент між обсягом вибірки в класі та кількістю помилок першого роду становить 0,759.

Аналогічною є залежність і для помилки другого роду.

Далі проаналізуємо якість кластеризації від кількості ключових слів, що її описує, а також від ступеню їх перетину. Алгоритм тестувався на чотирьох колекціях

вхідних даних з однаковою кількістю об'єктів у кожному з класів, але з різною кількістю ключових слів та з різною кількістю спільних для класів ключових слів.

Результати аналізу подано у таблицях 3 та 4.

Таблиця 3. Залежність якості кластеризації від кількості ключових слів

Клас	Колекція 1		Колекція 2	
	к-сть ключових слів	nTP	к-сть ключових слів	nTP
Database	7	87	16	88
computer science	11	67	26	62
programming	12	69	19	67
network	3	93	7	91
system analysis	4	94	5	89

Таблиця 4. Залежність якості кластеризації від кількості спільних для кластерів ключових слів

Клас	Колекція 3		Колекція 4	
	% спільних ключових слів	nTP	% спільних ключових слів	nTP
database	3	69	12	52
computer science	14	59	32	44
programming	13	61	27	47
network	5	68	14	51
system analysis	2	72	11	62

Як бачимо, якість кластеризації залежить більше від унікальності ключових слів у кластерах, а менше – від їх кількості.

Наступним кроком є визначення якості кластеризації для різних методів.

Для порівняння було проаналізовано результати роботи трьох інших алгоритмів на тих же колекціях.

Були отримані наступні результати (значення TP):

Таблиця 5. Порівняння результатів роботи різних методів кластеризації

Метод кластеризації	**nTP**
Розроблений метод кластеризації	0.92
Острівна кластеризація	0.86
К-середні	0.71
Average Link	0.78

Таким чином, розроблений алгоритм продемонстрував кращі результати у значенні величини nTp на текстових колекціях порівняно з іншими розглянутими алгоритмами.

Далі проаналізовано часову складність розробленого алгоритму кластеризації.

Усі методи тестувались на тому ж наборі даних і на тому ж комп’ютері: Intel Core 2 Quad E6600 2.4 GHz, 8 GB RAM, HDD WD 2 TB 7200 RPM. Для збереження даних використовувалась СКБД Microsoft SQL Server 2008 R2 Developer Edition.

Таблиця 6. Час аналізу (с) текстових об’єктів залежно від об’єму проаналізованих даних

Обсяг вибірки	Розроблений метод	Острівна кластеризація	К-середні	Average Link
20	9	8	9	10
50	12	11	13	13
100	15	15	16	15
150	18	17	21	21

Висновки

У роботі запропоновано метод виділення елементів для побудови текстового шаблону, а також метод кластеризації текстового шаблону.

Використання текстових шаблонів на основі ключових слів дозволяє опрацьовувати фактично будь-який слабоструктурований текст, якщо для нього складено словник ключових слів.

Зважаючи на результати дослідження, варто відзначити, що розроблений метод домінується лише методом острівної кластеризації.

Як бачимо, час виконання аналізу даних суттєво відрізняється для різних методів. Жоден метод на практиці не досягає лінійної складності алгоритму аналізу даних залежно від розміру набору даних, що аналізуються.

Література

1. Shakhovska, N.B., Noha, R.Y. 2015 . Methods and Tools for Text Analysis of Publications to Study the Functioning of Scientific Schools. Journal of Automation and Information Sciences, p. 47.
2. Захарчук Т.В. Научные школы в библиографоведении: особенности формирования / Т.В. Захарчук // Научно-техническая информация. Сер. 1. Организация и методика информационной работы.– 2011. – № 1. – С. 19–25.
3. Chappin E.J.L. Transition and transformation: A bibliometric analysis of two scientific networks researching socio-technical change / Emile J.L. Chappin, Andreas Ligtvoet // Renewable and Sustainable Energy Reviews. –2014. – Vol. 30. –P. 715–723.
4. Ланде Д.В. Наукометричні дослідження мереж співавторства по базі даних«Україніка наукова» / Д.В. Ланде, І.В. Балагура // Реєстрація, зберігання і обробка даних. – 2012, – Т.14, No4 –С.41-51.
5. Berry M., Kogan J. Text Mining. Applications and Theory. West Sussex: Wiley, 2010. - 222 p.
6. Park S.-T. Analysis of Lexical Signatures for Finding Lost or Related Documents / S.-T. Park, D. Pennock, C. Lee Giles, R. Krovetz. - Finland, 2002. - 8p.

Literatura

1. Shakhovska, N.B., Noha, R.Y. 2015 . Methods and Tools for Text Analysis of Publications to Study the Functioning of Scientific Schools. Journal of Automation and Information Sciences, p. 47.
2. Zaxarchuk T.V. Nauchnye shkoli v by`bly`ografovedeny`y`: osobennosty` formy`rovany`ya / T.V. Zaxarchuk // Nauchno-texny`ches-kaya y`nformacy`ya. Ser. 1. Organy`zacy`ya y` metody`ka y`nformacy`on-noj raboty.– 2011. – # 1. – S. 19–25.
3. Chappin E.J.L. Transition and transformation: A bibliometric analysis of two scientific networks researching socio-technical change / Emile J.L. Chappin, Andreas Ligtvoet // Renewable and Sustainable Energy Reviews. –2014. – Vol. 30.–P. 715–723.
4. Lande D.V. Naukometry`chni doslidzhennya merezh spivavtorstva po bazi dany`x «Ukrayinika naukova» / D.V. Lande, I.V. Balagura // Reyestraciya, zberigannya i obrobka dany`x. – 2012, – T.14, No4 –S.41-51.
5. Berry M., Kogan J. Text Mining. Applications and Theory. West Sussex: Wiley, 2010. - 222 p.
6. Park S.-T. Analysis of Lexical Signatures for Finding Lost or Related Documents / S.-T. Park, D. Pennock, C. Lee Giles, R. Krovetz. - Finland, 2002. - 8p.

RESUME

N.B. Shakhovska, I.B. Shvorob

Method of constructing a text template for extracting information from semistructured data

One of the main problems with semistructured data is the diversity of data. The question of the diversity of data in information systems is complex, it also includes areas such as data grouping for certain characteristics without sequential overlapping of sets of data.

To solve these issues, the paper suggests using the approach to constructing a text template based on the given keywords. This approach will help to remove excess information and obtain the necessary data, avoiding duplication.

A text template is an unstructured or semistructured file that consists of a sequence of sentences, and sentences from a sequence of words. Of the entire plural of words in the document, only those with content content are selected, that is, the "Keywords" database is formed.

In this paper, the main purpose is to develop a method for selecting component elements for constructing a text template, as well as developing a method for clustering a text template. The analysis of the developed methods on the example of work of the library system is carried out.

The result of the method of selecting the components of a text document is a vector in which for certain characteristics of the text used binary signs, and for keywords - weights. The method of clustering a text template uses an improved k-medium method.

Given the results of the study, it should be noted that the developed method is only dominated by the method of island clusterization. The time of data analysis is significantly different for different methods. No method in practice does not achieve the linear complexity of the data analysis algorithm, depending on the size of the data set being analyzed.

Надійшла до редакції 06.10.2017